Published online 28 October 2006 *Nucleic Acids Research, 2007, Vol. 35, Database issue* ***D575–D579***
doi:10.1093/nar/gkl768

PROTCOM: searchable database of protein complexes enhanced with domain–domain structures

Petras J. Kundrotas and Emil Alexov*

Computational Biophysics and Bioinformatics, Department of Physics and Astronomy, Clemson University, Clemson, SC 29634, USA

Received July 5, 2006; Revised September 13, 2006; Accepted October 1, 2006

ABSTRACT

The database of protein complexes (PROTCOM) is a compilation of known 3D structures of protein–protein complexes enriched with artificially created domain–domain structures using the available entries in the Protein Data Bank. The domain–domain structures are generated by parsing single chain structures into loosely connected domains and are important features of the database. The database (http://www.ces.clemson.edu/compbio/protcom) could be used for benchmarking purposes of the docking and other algorithms for predicting 3D structures of protein–protein complexes. The database can be utilized as a template database in the homology or threading methods for modeling the 3D structures of unknown protein–protein complexes. PROTCOM provides the scientific community with an integrated set of tools for browsing, searching, visualizing and downloading a pool of protein complexes. The user is given the option to select a subset of entries using a combination of up to 10 different criteria. As on July 2006 the database contains 1770 entries, each of which consists of the known 3D structures and additional relevant information that can be displayed either in text-only or in visual mode.

INTRODUCTION

The recent successes of the human and other genome projects provide the amino acid sequences of all macromolecules in the living cell. Utilizing this information will enable us to shed light on the mechanism of the cellular processes. Therefore, significant efforts are directed towards using the amino acid sequences to predict both the protein–protein interactions and the 3D structures of protein–protein complexes. Since experimental methods such as X-ray and NMR cannot be used at such large scales, developing computational methods that can predict 3D structures of protein–protein complexes is of great importance.

Currently, the 3D structures of protein–protein complexes are predicted by docking (1), binding simulations (2), fitting to electron microscopy data (3), threading (4) and homology modeling (5–7). The most popular method of predicting 3D structures of protein–protein complexes is the docking (1), which is broadly used in the Critical Assessment of Prediction of Interactions (CAPRI) competition (8–10). Recently, significant progress has also been made applying threading (4,11) and homology (6,7,12,13) methods to predict 3D structures of protein–protein complexes. Important aspect of the last two methods with respect to our database is that they both use templates to predict the 3D structures of protein–protein complexes. Thus, the databases such as iPfam (14), 3did (15), InterPare (16), PiBase (17) and PDBsum (18) that can provide templates for modeling protein–protein complexes are crucial for the success of the threading and homology approaches.

There are two different approaches for modeling 3D structures of complexes: full-length complex structure prediction (4) and domain–domain structure predictions (6). A database that provides templates for both approaches and has searching tools could be beneficial for both. The paper reports a database of protein–protein complexes enriched by domain–domain structures that were artificially created from single-chain structures in order to mimic protein–protein complexes. The inclusion of domain–domain structures in the form of two-chain 'complexes' increases almost 2-fold the pool of available structures and is an important feature of the PROTCOM database, which at present (as on July 2006) contains 1770 entries. The database is aimed to serve as a pool of templates for threading and homology modeling methods, as a benchmarking set of diverse structures of complexes for the docking and other methods of predicting 3D structures of protein–protein complexes, and also as a tool for investigating the properties of protein–protein complexes through its searching and visualization modules. In this respect PROTCOM offers unique features such as search for entries having (i) particular type of residue or a combination of residues present at the interface of the complex or any of the partners; (ii) particular interfacial

*To whom correspondence should be addressed. Tel: +1 864 656 5307; Fax: +1 864 656 0805; Email: ealexov@clemson.edu area; and (iii) particular X-ray resolution and any combination of the above. Another distinguishable feature of the PROTCOM database is that the original Protein Data Bank (PDB) files corresponding to single-chain or more than two-chain proteins are modified; hence, each entry in the database is made up of two interacting units.

DATABASE CONTENT AND DESCRIPTION

Selection principles

Current implementation of the PROTCOM database consists of two parts: two-chain protein–protein complexes (966 entries) and two-domain single-chain proteins (804 entries). The domain–domain entries are created from single-chain proteins that are split into domains using the PDP program (19). Only domains that are loosely connected by a single polypeptide chain that does not correspond to either helix or to strand are selected. The names of the PDB files of the domain–domain structures in the database are modified by adding the suffix '_com' after the PDB ID code. The content of the original PDB file is also modified by deleting two residues in the link between domains and assigning new chain designators, A and B, to the residues belonging to the first and second domains, respectively.

The proteins in the database are selected from X-ray structures available in the up-to-date local version of the PDB. All proteins in the selected set (both real two-chain complexes and two-domain structures) satisfy the following criteria:

(i) The sequence identity between any two entries should be <95%. This criterion is also applied to the sequences belonging to the same protein–protein complex since our primary goal is to create a database of heterocomplexes. However, an extension of the database that will include separate section of homocomplexes will be implemented in the new database release.

(ii) The area of the interface between A and B parts should be >250 $Å^2$ and <50% of surface accessible area of either component (hereafter, the notations A and B stand for the larger and smaller complex parts, respectively, which, in general, may not coincide with the experimental chain designators in the PDB file).

(iii) There should be at least two secondary structure elements (helix or strand) in each component of the complex.

Very often, X-ray PDB files of two-chain complexes contain several identical pairs, which are artifacts of crystallization. Such structures are also included in the PROTCOM database, but in this case, the original PDB files are modified so that only one pair out of several identical units is left. The experimental chain designators in those files and their names are left intact.

Structure of files

Files in the database are grouped into three folders: (i) the PDB folder with 3D structure in the PDB format, (ii) the SEQ folder with the sequences of both components in the FASTA format and (iii) the INF folder where all other relevant information is stored. File name nomenclature is a 4-symbol PDB code with extension 'pdb' (for the structure files), 'seq' (for the sequence file) or 'inf' (for the information files). Such organization of files is preserved when a user downloads search results (or the entire database).

Searching the database

The incorporated search engine (package of Javascript and Perl programs) offers search options in a very easy and user-friendly manner. The representative snapshot of the search page is shown in Figure 1. At the top of the page, the user is given an option to select within which part of the database (within the complexes only, the domain–domain structures, or the whole database) the search will be performed and what conjunction (AND or OR) will be used between the search criteria. If no further selection criteria are chosen, the search results will consist of the entire content of the corresponding parts of the database.

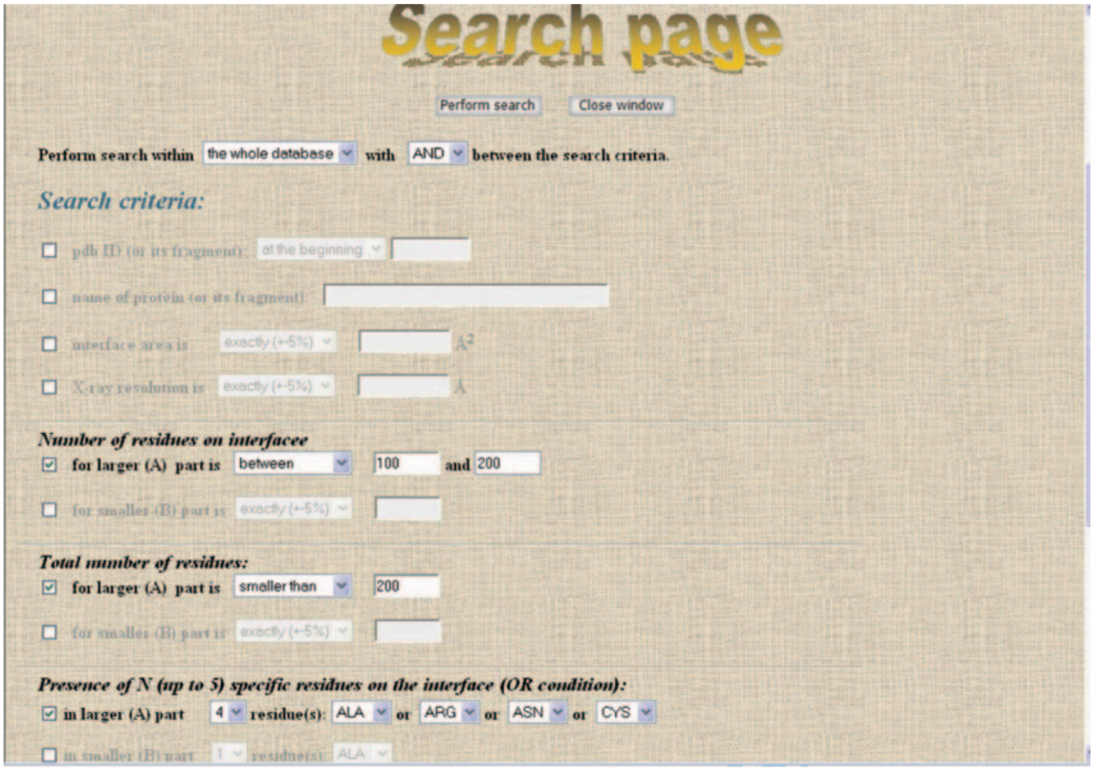
For better visual perception of the search screen, the name of a search criterion is blurred and the related fields are disabled when this criterion is not selected (see Figure 1). For the numerical inquiries, the user is given a choice to search for values exactly equal to (with 5% tolerance level), larger than, smaller than or between the desired value(s) (in the latter case, the additional field appears in the screen for upper threshold of the value). There are also options to search for the presence of up to five specific residues and for the presence of specific combination of up to five residues on the interface in the A and B parts, separately.

The search results can be sorted in ascending or descending order by a number of parameters and are presented as a table with a minimum set of relevant info, which includes PDB ID (with a link to the corresponding entry in the PDB), name of the protein, resolution of the X-ray structure, classification of the entry in the database (complex or domain–domain structure), interface area and number of residues (total and on the interface) in the larger A and smaller B parts of the entry. If a detailed analysis is desired, an option to check particular entries for further downloading is provided as well as links either to text-only or to visual displays of the full information about the particular structure.

Visualization of entries

The visualization screen (Figure 2) utilizes JMOL technology (www.jmol.org) and initially displays a ribbon representation of the secondary structure elements of both polypeptide chains along with the ball-and-stick representation of the interfacial residues. The ribbons are colored by the chain while the color scheme for the ball-and-stick models is determined by the type of atoms. Interfacial residues can be displayed using sticks-only, ball-and-sticks or spacefill (atoms are displayed as spheres with the radius equal to corresponding van der Waals radius) model representations. There is also a possibility to display either the van der Waals dot surface or the solvent accessible surface area for the interfacial residue.

For a more detailed examination of the interfacial residues, an option is given to view the whole complex or its separate components along with the possibility to view interfacial residues separately by checking corresponding boxes on the



Perform search within the whole database with AND between the search criteria.

Search criteria:

□ pdb ID (or its fragment): at the beginning

□ name of protein (or its fragment):

□ interface area is exactly (+-5%) Å²

□ X-ray resolution is exactly (+-5%) Å

Number of residues on interfacee

☑ for larger (A) part is between 100 and 200

□ for smaller (B) part is exactly (+-5%)

Total number of residues:

☑ for larger (A) part is smaller than 200

□ for smaller (B) part is exactly (+-5%)

Presence of N (up to 5) specific residues on the interface (OR condition):

☑ in larger (A) part 4 residue(s): ALA or ARG or ASN or CYS

□ in smaller (B) part 1 residue(s): ALA

Figure 1. Snapshot of the search page of the PROTCOM database.

bottom part of the visualization screen (Figure 2). The JMOL technology also provides other visualization options [for full description see the JMOL website (www.jmol.org)].

Access to the database

Access to the database (http://www.ces.clemson.edu/compbio/protcom) is free of charge and the users are given an option to send suggestions and comments to the authors of the database. Users can also freely download either a single entry or the entire database (or part of it) as a single 'tar.gz' file.

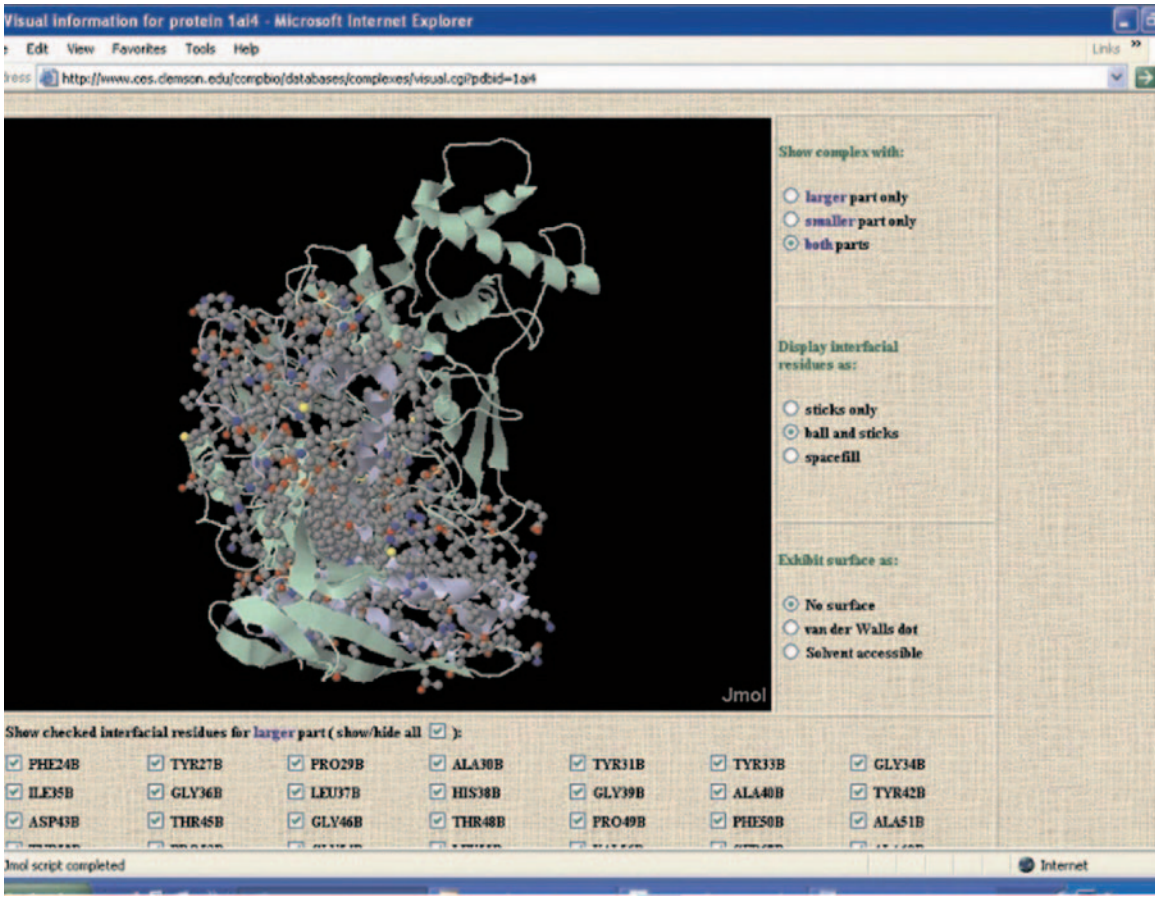
CURRENT STATE OF THE DATABASE

As on July 2006, the database contains 1770 entries of which 966 are the two-chain protein–protein complexes and 804 domain–domain structures. In the current release of the PROTCOM database each entry accommodates the following information: PDB ID, the name of the protein, the resolution of the X-ray structure, the calculated absolute interface area, the total number of residues, the number of residues on the interface, the relative interface area, the list of the interfacial residues, and the number of helices and strands (separately for A and B parts). A residue was included in the list of interfacial residues if the distance between any heavy atom of the residue and any heavy atom of any residue belonging to another component of the structure is less than the sum of van der Waals radii of the atoms plus the diameter of water molecule, 2.8 Å. For informational purposes the sequences of both components of the database entry in the FASTA format are also displayed. The search results are arranged by the PDB ID, structure name, X-ray resolution, absolute interface area, total number of residues in the A and B parts, and the total number of interfacial residues in the A and B parts. In addition, the user can search for the presence of up to five specific residues on the interface in the A and B parts and for the presence of specific combination of up to five residues.

We have also developed an update script which automatically runs each month in order to create an updated version of the PROTCOM database based on the local copy of the PDB which, in turn, is automatically updated on a weekly basis.

COMPARISON TO OTHER DATABASES

As it was mentioned in Introduction, currently there are several databases of protein complexes and protein–protein interfaces. In order to provide the user with an opportunity to access the alternative resources, links to other relevant databases are added on the PROTCOM front page. In comparison to other databases, the PROTCOM database possesses a number of distinctive features with respect to both the content and the functional capabilities. From a structural perspective, the existing domain–domain databases iPfam (14), 3DID (15), PIBASE (17) and InterPare (16) were created with a purpose to investigate protein interactions on the domain level and thus do not include full-length protein complexes in the way PROTCOM does. The domain boundaries were defined using either the Pfam (20,21), SCOP (22) or CATH (23) definitions. The domain boundaries of the domain–domain entries in the PROTCOM are defined


Visual information for protein 1ai4 - Microsoft Internet Explorer
Edit View Favorites Tools Help
Links »
http://www.ces.clemson.edu/compbio/databases/complexes/visual.cgi?pdbid=1ai4
Jmol
Jmol script completed
Internet

Show complex with:

○ larger part only
○ smaller part only
● both parts

Display interfacial residues as:

○ sticks only
● ball and sticks
○ spacefill

Exhibit surface as:

● No surface
○ van der Walls dot
○ Solvent accessible

Show checked interfacial residues for larger part (show/hide all ☑):

☑ PHE24B	☑ TYR27B	☑ PRO29B	☑ ALA30B	☑ TYR31B	☑ TYR33B	☑ GLY34B
☑ ILE35B	☑ GLY36B	☑ LEU37B	☑ HIS38B	☑ GLY39B	☑ ALA40B	☑ TYR42B
☑ ASP43B	☑ THR45B	☑ GLY46B	☑ THR48B	☑ PRO49B	☑ PHE50B	☑ ALA51B

Figure 2. Visual display of information for the 1a0p entry of the PROTCOM database.

using PDP (19) domain parser and thus provide an alternative to the above classifications. In addition, domain–domain entries in the PROTCOM are artificially broken into two independent units in contrast to other domain–domain databases where the domains are continuous segments within the same polypeptide chain. Another related database is the PDBsum database (18), which provides an overview of every macromolecular structure deposited in the PDB. However, this database does not contain domain–domain structures and has many redundant entries. Several databases (such as the SPIN-PP database developed in Honig lab) and servers [such as the popular ConSurf (24,25) server] focus on the important aspects of the interfacial properties of protein–protein complexes. These resources will be used in the next generation of the PROTCOM database.

From a functional perspective, the PROTCOM database offers an option to create a subset of the database content with respect to user-defined combination of up to 10 different parameters (see Searching the database subsection) including searching for the presence of specific residues or a combination of residues at the interface, which we believe is a unique PROTCOM feature. Furthermore, the downloadable files are in pre-compiled format, and from the point of view of the user, even the domain–domain structures look like two-chain protein complexes and can directly be used in whatever software the user has for calculating properties of protein–protein complexes.

FUTURE DIRECTIONS

We plan for several directions of expanding and improving our database. We also plan to include in the PROTCOM database, complexes with several (>3) non-identical chains as well as structures parsed into several (>3) domains. From a functional perspective, the future version of the PROTCOM will give the user options to submit the selected entry to the SPIN-PP database and ConSurf server for further investigation of the interfacial properties. In addition, the database structure allows easy incorporation of new information and search parameters, which will undoubtedly arise from the feedback of the database users.

ACKNOWLEDGEMENTS

The authors thank Barry Honig for the continuous support. E.A. acknowledges startup funds from the Clemson University. The open access publication charges for this article were waivered by Oxford University Press.

Conflict of interest statement. None declared.

REFERENCES

1. Smith,G.R. and Sternberg,M.J.E. (2002) Prediction of protein–protein interactions by docking methods. *Curr. Opin. Struct. Biol.*, **12**, 28–35.

2. McCammon,J.A. (1998) Theory of biomolecular recognition. *Curr. Opin. Struct. Biol.*, **8**, 245–249.
3. Topf,M. and Sali,A. (2005) Combining electron microscopy and comparative modeling. *Curr. Opin. Chem. Biol.*, **15**, 578–585.
4. Lu,L., Lu,H. and Skolnick,J. (2005) MULTIPROSPECTOR: an algorithm for the prediction of protein–protein interactions by multimeric threading. *Proteins*, **15**, 350–364.
5. Aloy,P. and Russell,R.B. (2003) InterPreTS: protein interaction prediction through tertiary structure. *Bioinformatics*, **19**, 161–162.
6. Aloy,P. and Russell,R.B. (2006) Structural systems biology: modelling protein interactions. *Nature Rev. Mol. Cell Biol.*, **7**, 188–197.
7. Kundrotas,P.J. and Alexov,E. (2006) Predicting 3D structures of transient protein–protein complexes by homology. *Biochim. Biophys. Acta*, **1764**, 1498–1511.
8. Mendez,R., Leplae,R., Lensink,M. and Wodak,S. (2005) Assessment of CAPRI predictions in rounds 3–5 shows progress in docking procedures. *Proteins*, **60**, 150–169.
9. Vajda,S., Vakser,I., Steinberg,M. and Janin,J. (2002) Modeling of protein interactions in genomes. *Proteins*, **47**, 444–446.
10. Janin,J. (2005) The targets of CAPRI rounds 3–5. *Proteins*, **60**, 170–175.
11. Szilagyi,A., Grimm,V., Arakaki,A. and Skolnick,J. (2005) Prediction of physical protein–protein interactions. *Phys. Biol.*, **2**, 1–16.
12. Aloy,P., Ciccarelli,F.D., Leutwein,C., Gavin,A.C., Superti-Furga,G., Bork,P., Bottcher,B. and Russell,R.B. (2002) A complex prediction: three-dimensional model of the yeast exosome. *EMBO Rep.*, **3**, 628–635.
13. Aloy,P., Bottcher,B., Ceulemans,H., Leutwein,C., Mellwig,C., Fischer,S., Gavin,A.C., Bork,P., Superti-Furga,G., Serrano,L. *et al.* (2004) Structure-based assembly of protein complexes in yeast. *Science*, **303**, 2026–2029.
14. Finn,R.D., Marshall,M. and Bateman,A. (2005) iPfam: visualization of protein–protein interactions in PDB at domain and amino acid resolutions. *Bioinformatics*, **21**, 410–412.
15. Stein,A., Russell,R.B. and Aloy,P. (2005) 3did: interacting protein domains of known three-dimensional structure. *Nucleic Acids Res.*, **33**, D413–D417.
16. Gong,S., Park,C., Choi,H., Ko,J., Jang,I., Lee,J., Bolser,D.M., Oh,D., Kim,D.S. and Bhak,J. (2005) A protein domain interaction interface database: InterPare. *BMC Bioinformatics*, **6**, 207.
17. Davis,F.P. and Sali,A. (2005) PIBASE: a comprehensive database of structurally defined protein interfaces. *Bioinformatics*, **21**, 1901–1907.
18. Laskowski,R.A., Chistyakov,V.V. and Thornton,J.M. (2005) PDBsum more: new summaries and analyses of the known 3D structures of proteins and nucleic acids. *Nucleic Acids Res.*, **33**, D266–D268.
19. Alexandrov,N. and Shindyalov,I. (2003) PDP: protein domain parser. *Bioinformatics*, **19**, 429–430.
20. Bateman,A., Coin,L., Durbin,R., Finn,R.D., Hollich,V., Griffiths-Jones,S., Khanna,A., Marshall,M., Moxon,S., Sonnhammer,E.L. *et al.* (2004) The Pfam protein families database. *Nucleic Acids Res.*, **32**, D138–D141.
21. Finn,R.D., Mistry,J., Schuster-Bockler,B., Griffiths-Jones,S., Hollich,V., Lassmann,T., Moxon,S., Marshall,M., Khanna,A., Durbin,R. *et al.* (2006) Pfam: clans, web tools and services. *Nucleic Acids Res.*, **34**, D247–D251.
22. Andreeva,A., Howorth,D., Brenner,S.E., Hubbard,T.J., Chothia,C. and Murzin,A.G. (2004) SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Res.*, **32**, D226–D229.
23. Orengo,C.A., Pearl,F.M. and Thornton,J.M. (2003) The CATH domain structure database. *Methods Biochem. Anal.*, **44**, 249–271.
24. Glaser,F., Steinberg,D., Vakser,I. and Ben-Tal,N. (2001) Residue frequencies and pairing preferences at protein–protein interfaces. *Proteins*, **43**, 89–102.
25. Landau,M., Mayrose,I., Rosenberg,Y., Glaser,F., Martz,E., Pupko,T. and Ben-Tal,N. (2005) ConSurf 2005: the projection of evolutionary conservation scores of residues on protein structures. *Nucleic Acids Res.*, **33**, W299–W302.